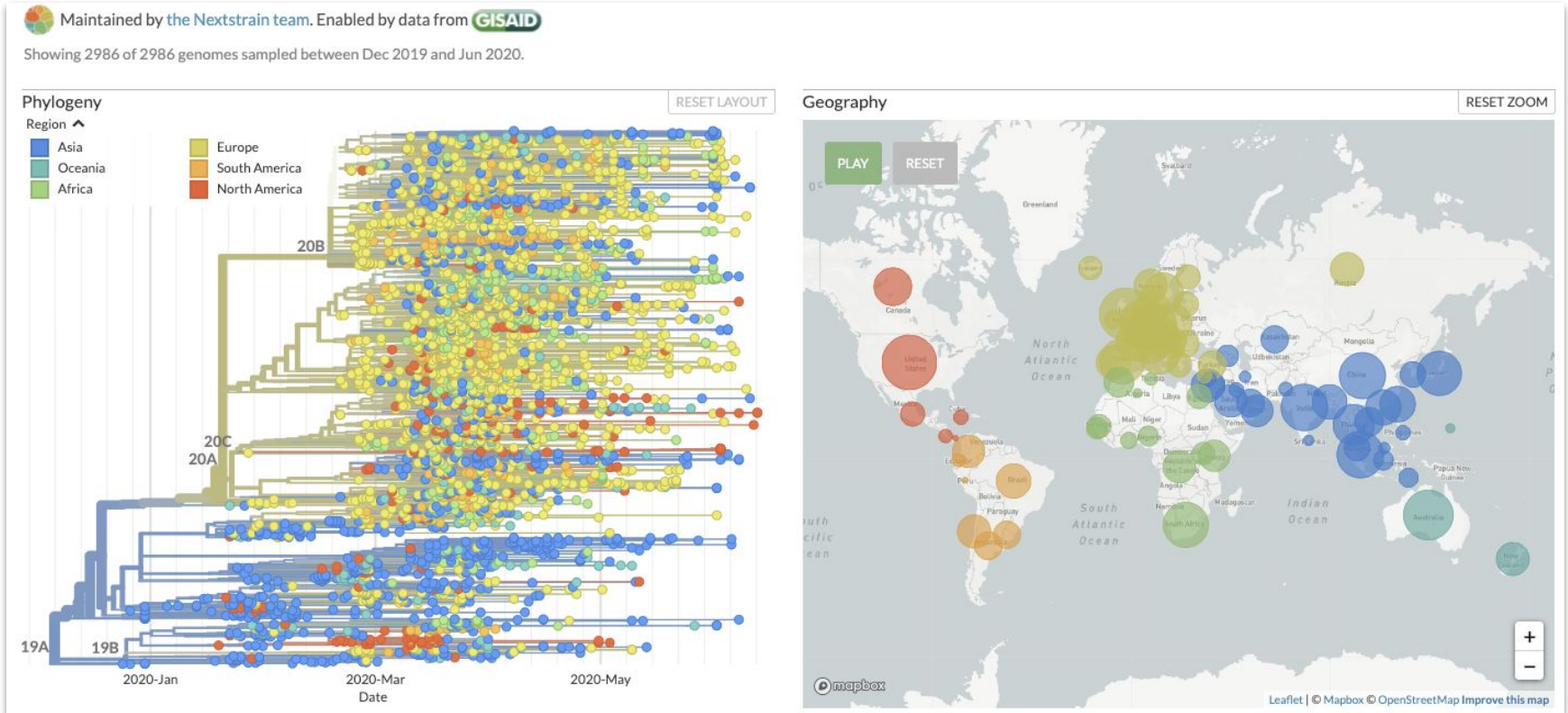


Building and analyzing SARS-CoV-2 consensus genomes



Consensus genomes are necessary!

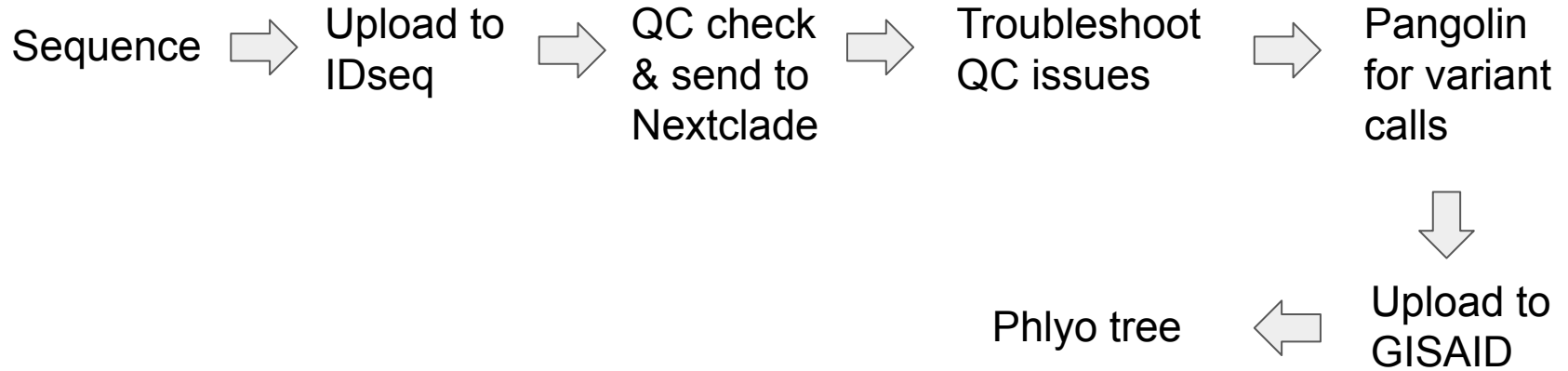
In order to make the trees to interpret transmission, you need to build consensus genomes



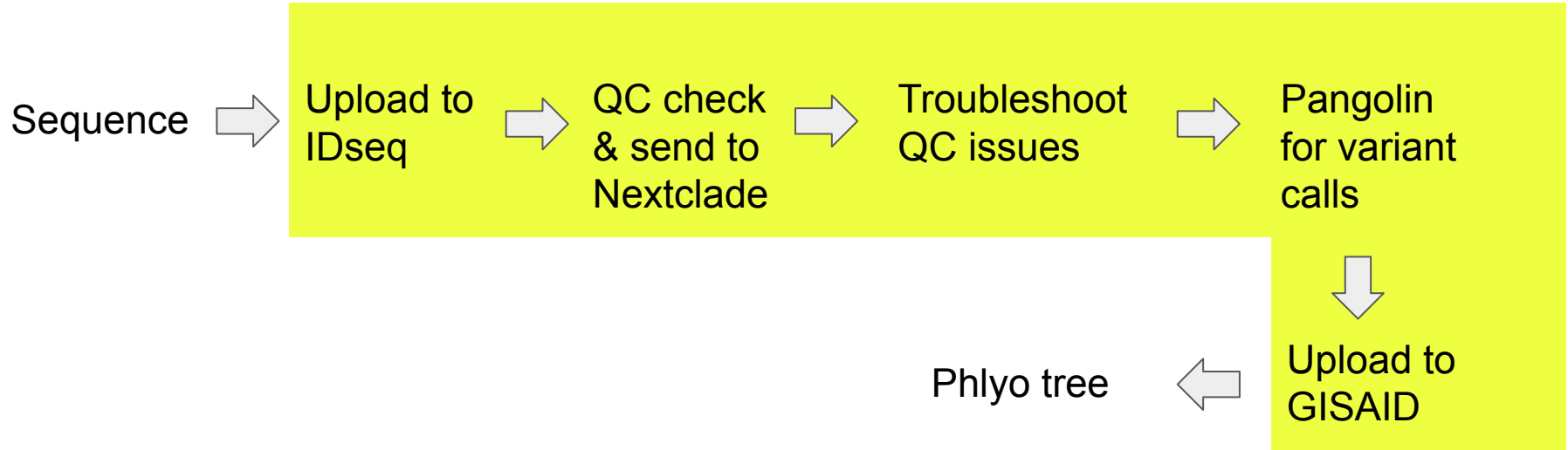
Consensus genome- represents multiple aligned reads



Workflow overview

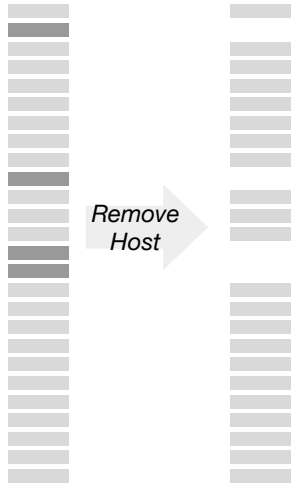


Workflow overview



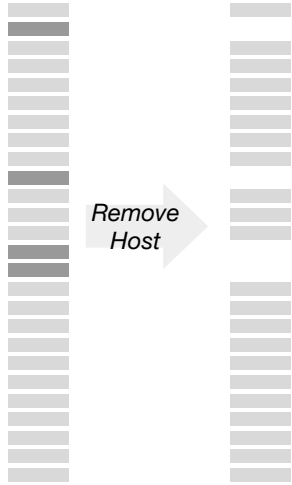
Generating Consensus Genomes

Raw reads



Generating Consensus Genomes

Raw reads



Remove
Host

Alignment

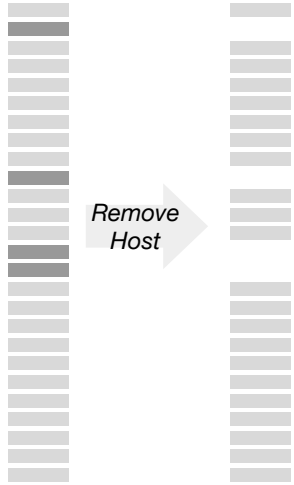
Reads mapped to reference



SARS-CoV-2 Reference Genome

Generating Consensus Genomes

Raw reads



Remove
Host

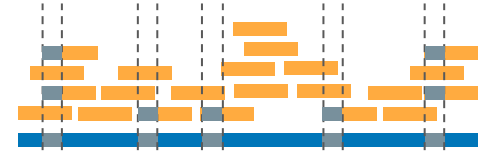
Alignment

Reads mapped to reference



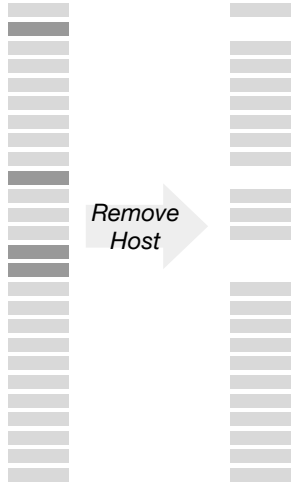
SARS-CoV-2 Reference Genome

Trim
Primers



Generating Consensus Genomes

Raw reads



Remove
Host

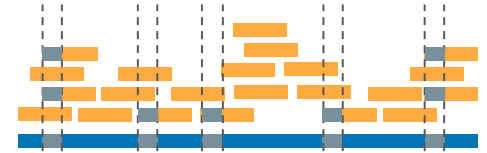
Alignment

Reads mapped to reference



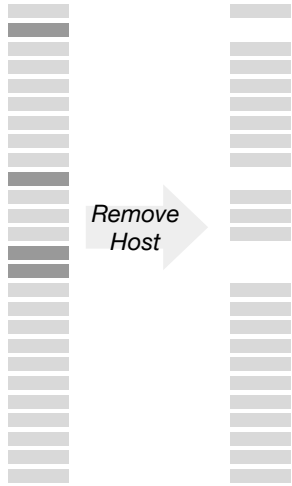
SARS-CoV-2 Reference Genome

Trim
Primers



Generating Consensus Genomes

Raw reads



Remove Host

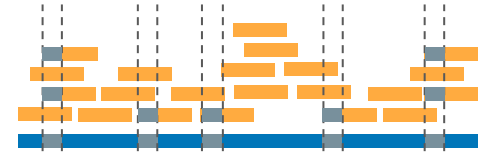
Alignment

Reads mapped to reference



SARS-CoV-2 Reference Genome

Trim Primers



Consensus Sequence

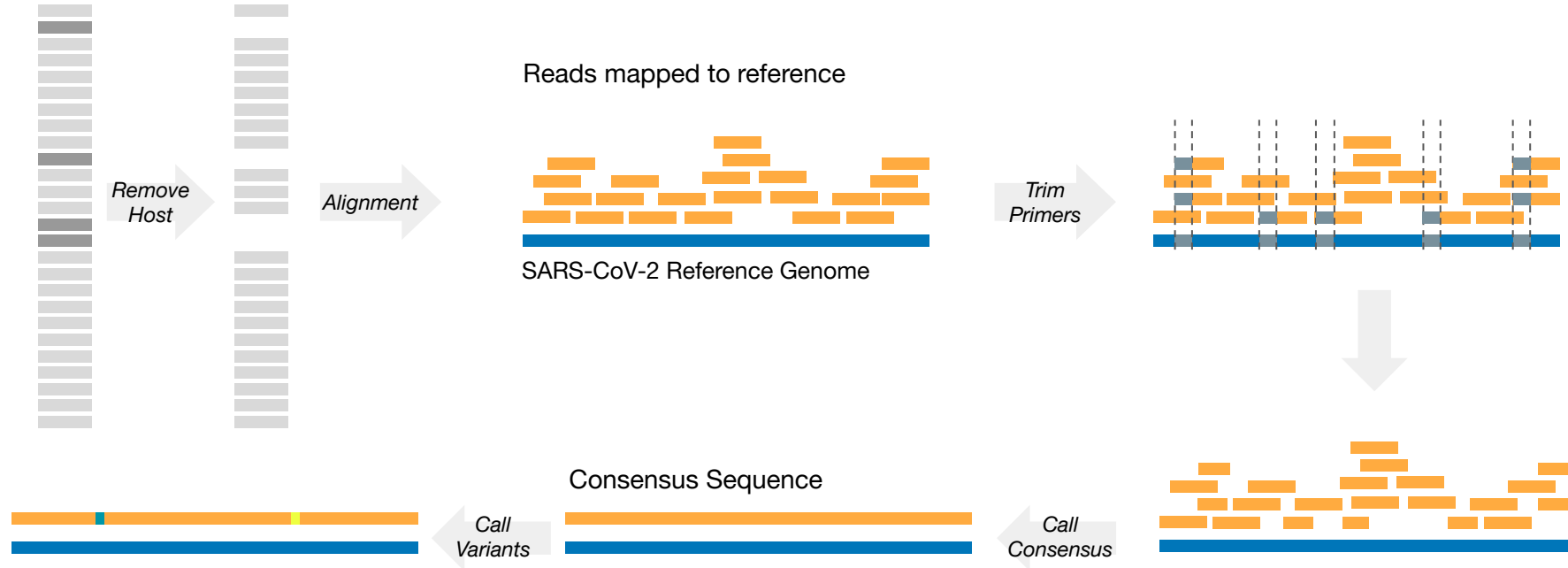


Call Consensus

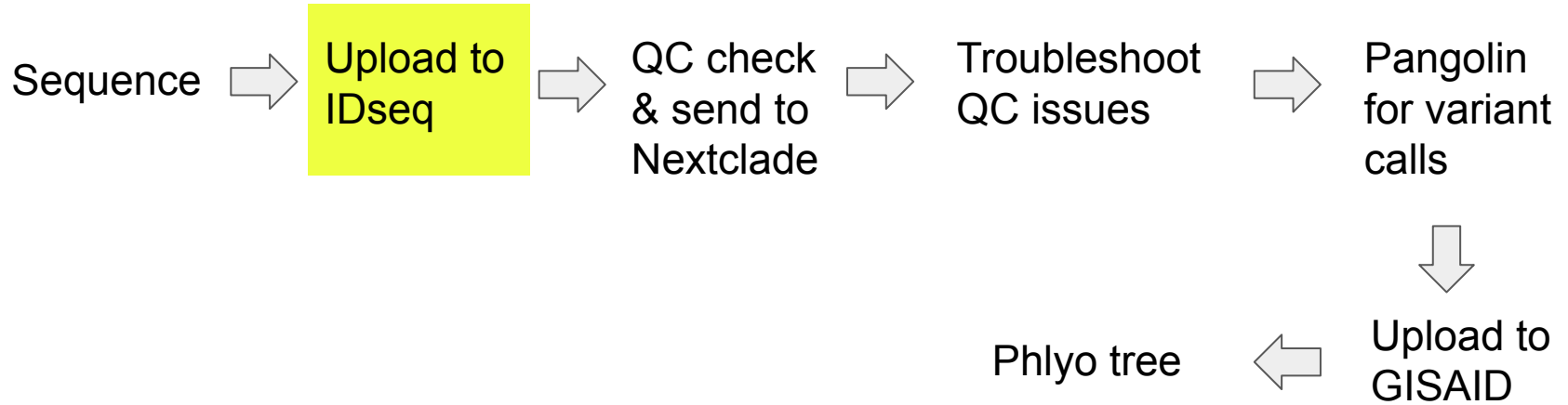


Generating Consensus Genomes

Raw reads



Workflow overview



Supports Illumina and Nanopore platforms

Select Project


Project


Select project



[+ CREATE PROJECT](#)

Analysis Type

 **Metagenomics**
Run your samples through our metagenomics pipeline. Our pipeline only supports Illumina.

 **SARS-CoV-2 Consensus Genome**
Run your samples through our Illumina or Nanopore supported pipelines to get consensus genomes for SARS-CoV-2.

Sequencing Platform:

Illumina
You can check out the Illumina pipeline on GitHub [here](#).

Nanopore
We are using the ARTIC network's nCoV-2019 novel coronavirus bioinformatics protocol for nanopore sequencing, which can be found [here](#).

Upload Files

[Upload from Your Computer](#) [Upload from Basespace](#)

Upload Your Input Files [MORE INFO](#)

Drag and drop your files here, or [click to use a file browser](#).

Add metadata

Upload Metadata

This metadata will provide context around your samples and results in IDseq.

1

Samples

2

Metadata

3

Review

Required fields: We require the following metadata to determine how to process your data and display the results: Host Organism, Sample Type, Water Control, Nucleotide Type, Collection Date, Collection Location. Please be as accurate as possible! [View Full Metadata Dictionary](#).

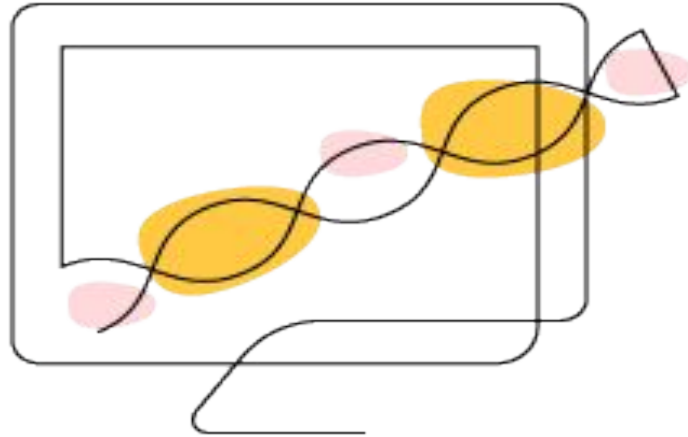
Available organisms for host subtraction: Human, Mosquito, Tick, Mouse, Cat, Pig, C.elegans, Carp, Chicken, Bee, Salpingoeca rosetta, Bat, Rat, Field Vole, Bank Vole, Rabbit, Water Buffalo, Horse, Taurine Cattle, Turkey, Barred Hamlet, Orange Clownfish, Tiger Tail Seahorse, Torafugu, Avian, White Shrimp.

Manual Input

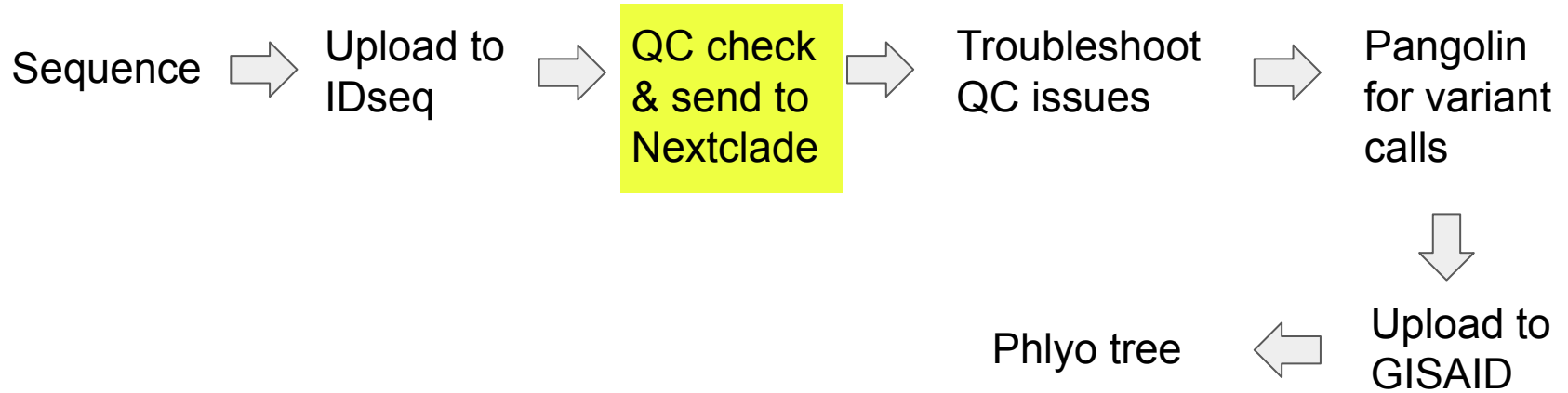
CSV Upload

Sample Name	Host Organism	Sample Type	Water Control	Nucleotide Type	Collection Date	Collection Location	
upload_file	<input type="text" value="v"/>	<input type="text" value="v"/>	<input type="radio"/> No	<input type="text" value="v"/>	<input type="text" value="YYYY-MM-DD"/>	<input type="text" value="Enter a city, region or country"/>	<input type="text" value="Q"/>

Pipeline runs automatically in the cloud



Workflow overview



Quality control check

CI >

sample1_2 ▾

[Sample Details](#)

[Download All](#)

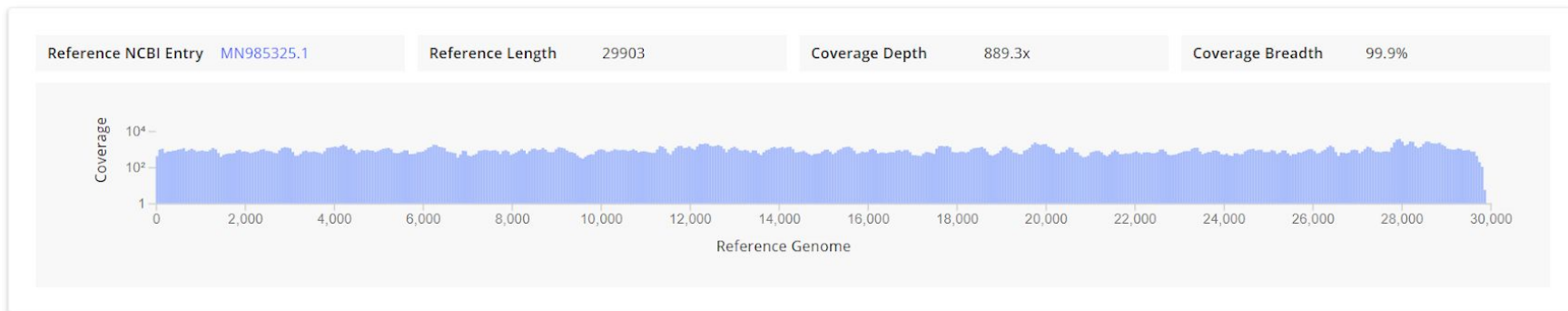
[Consensus Genome](#) BETA

[Learn more about consensus genomes >](#)

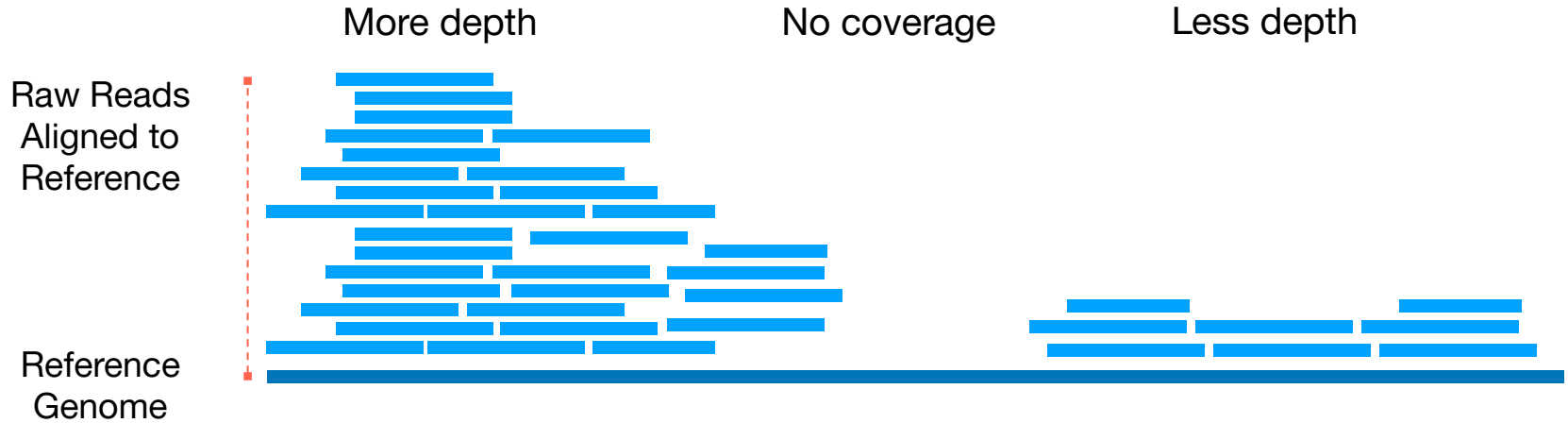
Is my consensus genome complete? ⓘ

Taxon	Reads	GC Content	SNPs	%id	Informative Nucleotides	Missing Bases	Ambiguous Bases
Severe acute respiratory syndrome coronavirus 2	187444	38.01%	7	100%	29850	12	0

How good is the coverage? ⓘ



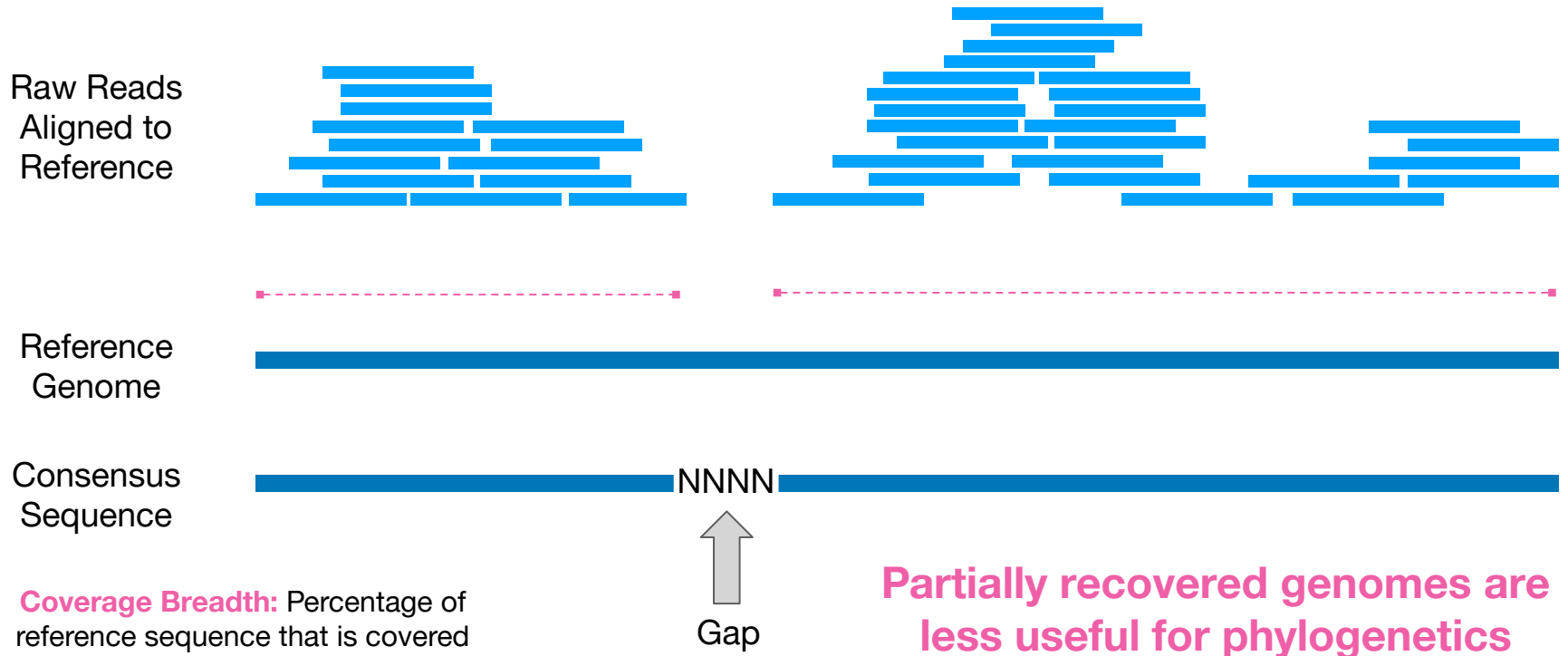
Is there enough depth?



Coverage Depth: # of times a nucleotide is read during sequencing

Must have >10 reads to call a base

How much of the genome was recovered?



How many SNPs are too many?



SNPs: Single Nucleotide Polymorphisms, variations in a single base pair in a DNA sequence

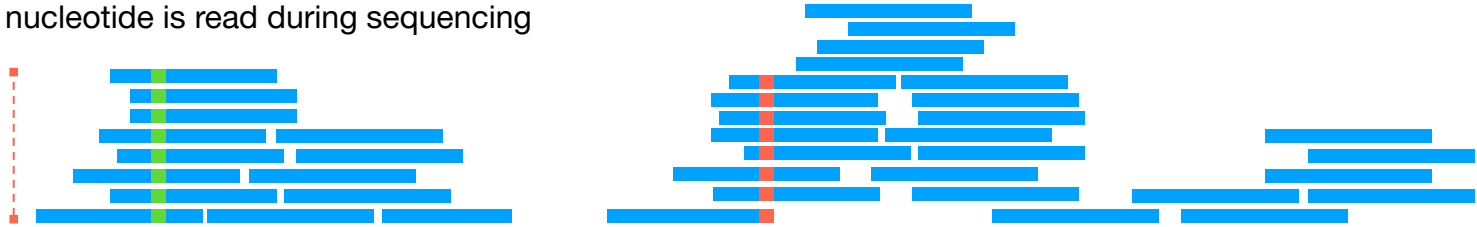
Pathogen dependent

- SARS-CoV-2 mutates slowly and shouldn't have too many SNP differences
 - Red flag: 30 differences. Orange flag: 25. Yellow flag: 20.

Evaluating Consensus Genomes

Coverage Depth: # of times a nucleotide is read during sequencing

Raw Reads
Aligned to
Reference



Coverage Breadth: Percentage of reference sequence that is covered

Reference
Genome



C

A

Consensus
Sequence



G

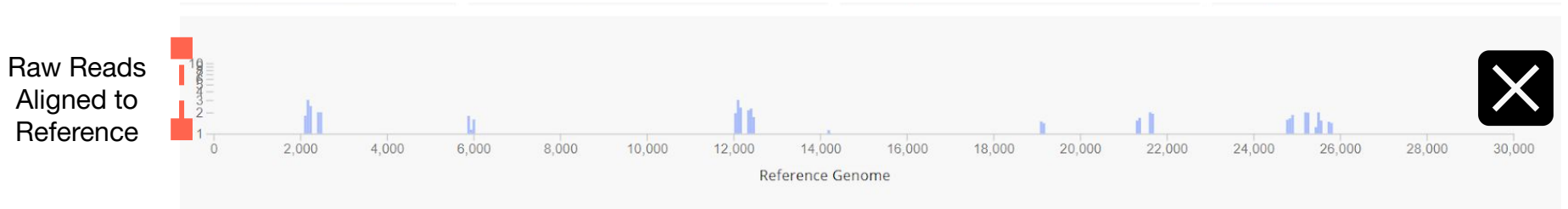
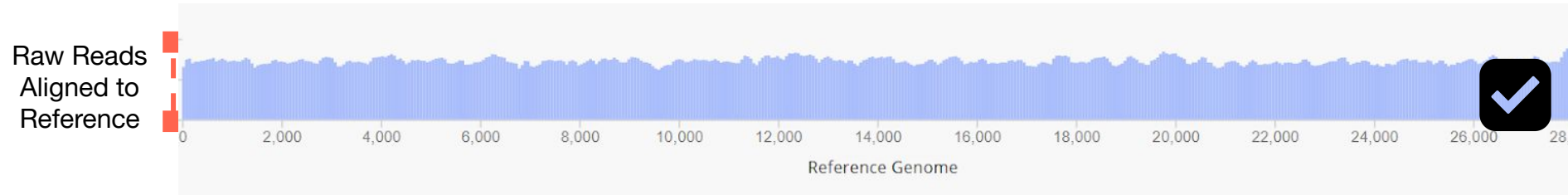
C

SNPs: Single Nucleotide Polymorphisms, variations in a single base pair in a DNA sequence

The coverage plot is a great first QC check

Coverage Depth: # of times a nucleotide is read during sequencing

Must have >10 reads in a location for a base to be called



Important metrics associated with the CG

CI >

sample1_2 ▾

[Sample Details](#)

[Download All](#)

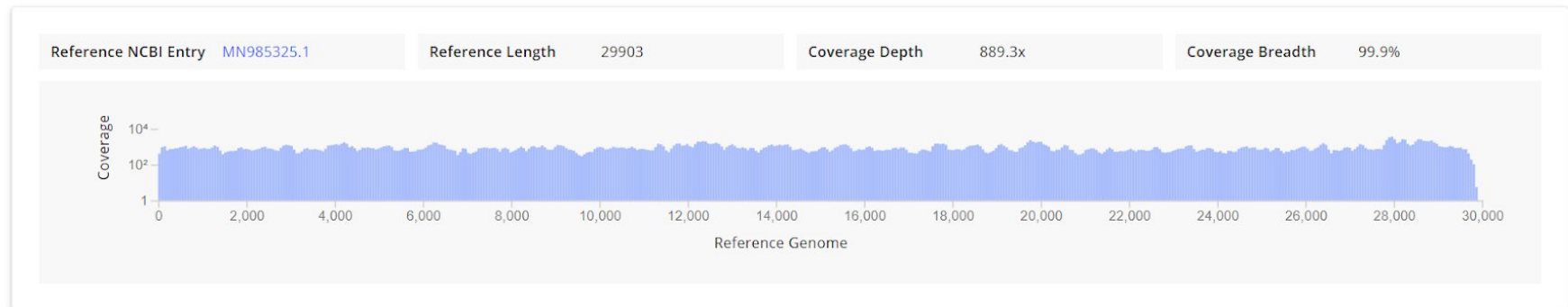
[Consensus Genome](#) BETA

[Learn more about consensus genomes >](#)

Is my consensus genome complete? ⓘ

Taxon	Reads	GC Content	SNPs	%id	Informative Nucleotides	Missing Bases	Ambiguous Bases
Severe acute respiratory syndrome coronavirus 2	187444	38.01%	7	100%	29850	12	0

How good is the coverage? ⓘ



Important metrics associated with the CG

CI >

sample1_2 ▾

[Sample Details](#)

[Download All](#)

[Consensus Genome](#) BETA

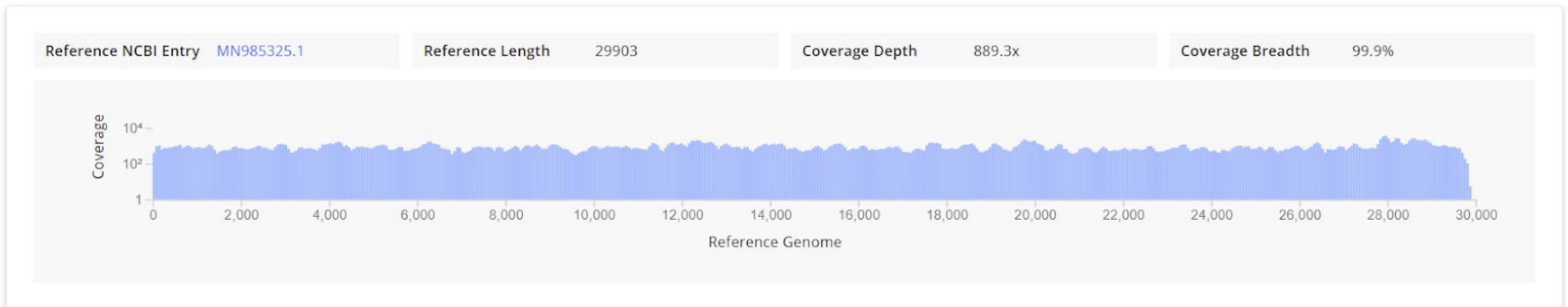
>30 

[Learn more about consensus genomes >](#)

Is my consensus genome complete? ⓘ

Taxon	Reads	GC Content	SNPs	%id	Informative Nucleotides	Missing Bases	Ambiguous Bases
Severe acute respiratory syndrome coronavirus 2	187444	38.01%	7	100%	29850	12	0

How good is the coverage? ⓘ



Important metrics associated with the CG

CI >

sample1_2 ▾

[Sample Details](#)

[Download All](#)

Nextstrain
requires
92% of ref
genome
coverage
(>27,510)

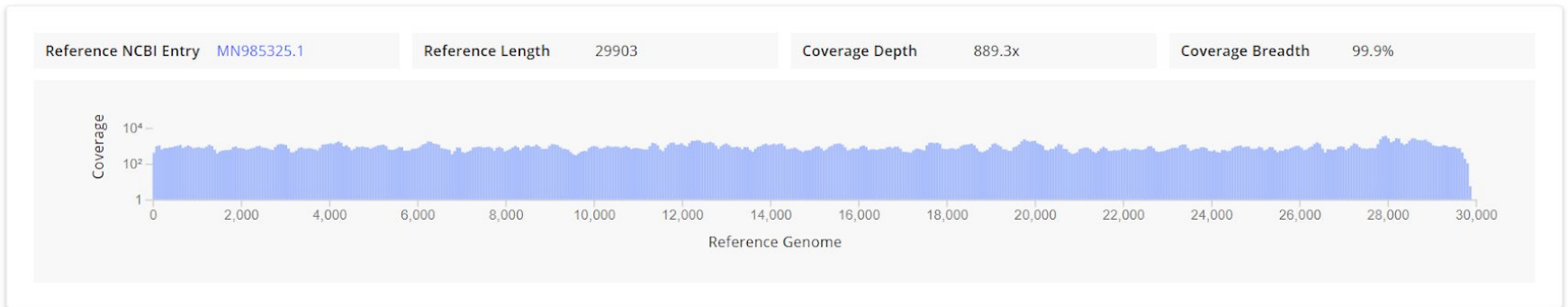
[Learn more about consensus genomes >](#)

[Consensus Genome](#) BETA

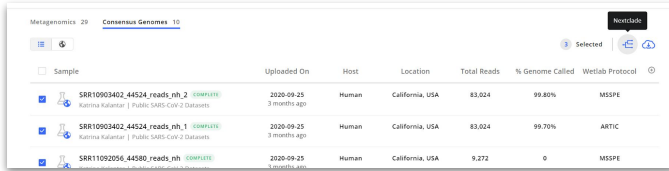
Is my consensus genome complete? ⓘ

Taxon	Reads	GC Content	SNPs	%id	Informative Nucleotides	Missing Bases	Ambiguous Bases
Severe acute respiratory syndrome coronavirus 2	187444	38.01%	7	100%	29850	12	0

How good is the coverage? ⓘ

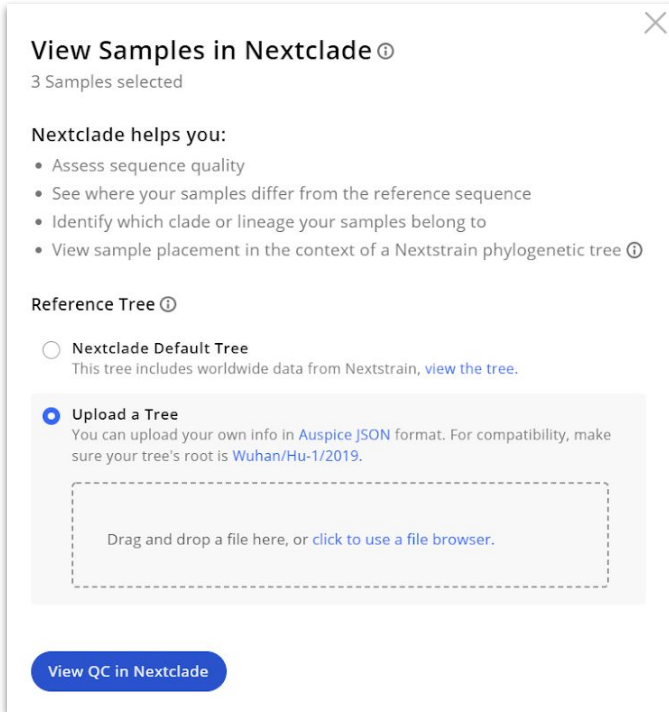


Send samples directly to Nextclade



Metagenomics 29 Consensus Genomes 10

Sample	Uploaded On	Host	Location	Total Reads	% Genome Called	Wetlab Protocol
SRR11092502.44524_reads_rh_2 <small>CONVITE</small> <small>Katrina Subtotal Public Health CDR 2 Datasets</small>	2020-09-25 <small>3 months ago</small>	Human	California, USA	83,024	99.80%	MSPF
SRR11092502.44524_reads_rh_1 <small>CONVITE</small> <small>Katrina Subtotal Public Health CDR 2 Datasets</small>	2020-09-25 <small>3 months ago</small>	Human	California, USA	83,024	99.70%	ARTIC
SRR11092505.44590_reads_rh <small>CONVITE</small> <small>Katrina Subtotal Public Health CDR 2 Datasets</small>	2020-09-25 <small>3 months ago</small>	Human	California, USA	9,272	0	MSPF



View Samples in Nextclade ⓘ
3 Samples selected

Nextclade helps you:

- Assess sequence quality
- See where your samples differ from the reference sequence
- Identify which clade or lineage your samples belong to
- View sample placement in the context of a Nextstrain phylogenetic tree ⓘ

Reference Tree ⓘ

Nextclade Default Tree
This tree includes worldwide data from Nextstrain, [view the tree](#).

Upload a Tree
You can upload your own info in [Auspice JSON](#) format. For compatibility, make sure your tree's root is [Wuhan/Hu-1/2019](#).

Drag and drop a file here, or [click to use a file browser](#).

[View QC in Nextclade](#)

- [Further investigate the quality of your consensus genomes in Nextclade.](#)
- Identify which clade or lineage your sample belongs to.
- Upload an existing tree or use the Nextclade default tree.
- Export auspice.json file from Nextclade.
- View phylogenetic tree with sensitive in a safe and secure environment ([Auspice](#)).

Nextclade results

- Sent samples >92% genome coverage (Nextstrain requires this to be added to their builds)



Settings

What's new

English



Back

Done. Total sequences: 8. Succeeded: 8



ID	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Gaps	Ins.	Nucleotide sequence
0	✓ CASC20008_L001	N M P C F S	20B	13	1	274	0	0	
1	✓ CASC20011_L001	N M P C F S	20A	14	1	280	0	0	
2	✓ SRR10903402_44524_reads_nh_1	N M P C F S	19A	0	4	60	0	0	
3	✓ SRR10971381_44496_reads_nh	N M P C F S	19A	0	115	3757	0	0	
4	✓ SRR10903402_44524_reads_nh	N M P C F S	19A	0	2	45	0	0	
5	✓ SRR10903401_44525_reads_nh	N M P C F S	19A	0	28	1022	0	0	
6	⚠ unknown_S1_L001	N M P C F S	19B	4	6	400	4	0	
7	✓ sample_sars-cov-2_paired	N M P C F S	20C	7	0	12	0	0	

Nextclade: Phylogenetic-based sequence QC

N

Number of sites where a base could not be called: Areas with low or no sequencing coverage have no information to tell you which base should be at that site. These sites are labelled with N's. When a sequence has too many N's it is both hard to align and place on the tree, and thus they are removed from analyses. By default Nextstrain will drop sequences with less than 27,000 non ambiguous bases.

M

Mixed sites: If many sequencing reads support *more* than one base at a site, those sites will be designated with an IUPAC ambiguity code, that tells you which *set* of mutations were found at the site. While this can happen given a co-infection event, it more commonly occurs due to sample cross-contamination.

P

Private mutations: If a sequence differs from the Wuhan reference genome by (currently) more than 20 mutations, it will be flagged as having a high number of “private” mutations. The threshold for flagging a sequence as problematic *will be changed* as the diversity of SARS-CoV-2 increases over the pandemic.

Nextclade: Phylogenetic-based sequence QC



C

Clusters of mutations: If your sequence has one or more areas with 6 mutations within a 100nt wide window, then that will be considered a “cluster of mutations” and it will be flagged unless it occurs at a recognized area of the genome. Such clusters of mutations are often artefactual, resulting from challenges aligning the sequence.



S

The presence of premature stop codons: a stop codon within a gene will now result in a QC warning, unless it is one of the very common stop codons in ORF8 at positions 27 or 68. Depending on where it is, it can be the result of an erroneous mutation.



F

The presence of frameshift mutations: This happens when there is an insertion or deletion that causes a gene to have a length that is not divisible by 3. If at least one such gene length is detected, the check is considered "bad". Failure of this check means that the gene likely fails to translate.

Nextclade: Phylogenetic-based sequence QC, in pictures

N CZI - AGTTCC**NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN**GTAAAC

M CZI - ATG**R**GAGTAAC**M**GGT**RW**TTT**G**ACCAGACACAC**A**M**G**ATT**BD**GGGA

P wuhan1 - AGTT**G**GTCC**A**TGATT**C**GTT**T**CGT**A**AATTCGT**C**TT**C**G**A**CAGTT**G**GT
CZI - AGTT**C**GTCC**T**TGATT**G**GTT**A**CGT**T**AATTCGT**G**TT**C**G**T**GAGTT**C**GT

Nextclade: Phylogenetic-based sequence QC, in pictures



wuhan1 - AGTTGGTCCATGATTCGTTTCGTTAATTCGTCTTCGACAGTTGGT
CZI - AGTTGGTCC**TACGGTGGTTAGAAA**TTTTTCGT**GTACCAGAGTTCGT**



wuhan1 - AGTTGGTCCATGATTCGTTTCGTCTATTCGTCTTCGACAGT**TAA**
CZI - AGTTGGTCCATGATTCGTTTCGT**TAA**ATTCGTCTTCGACAGT**TAA**



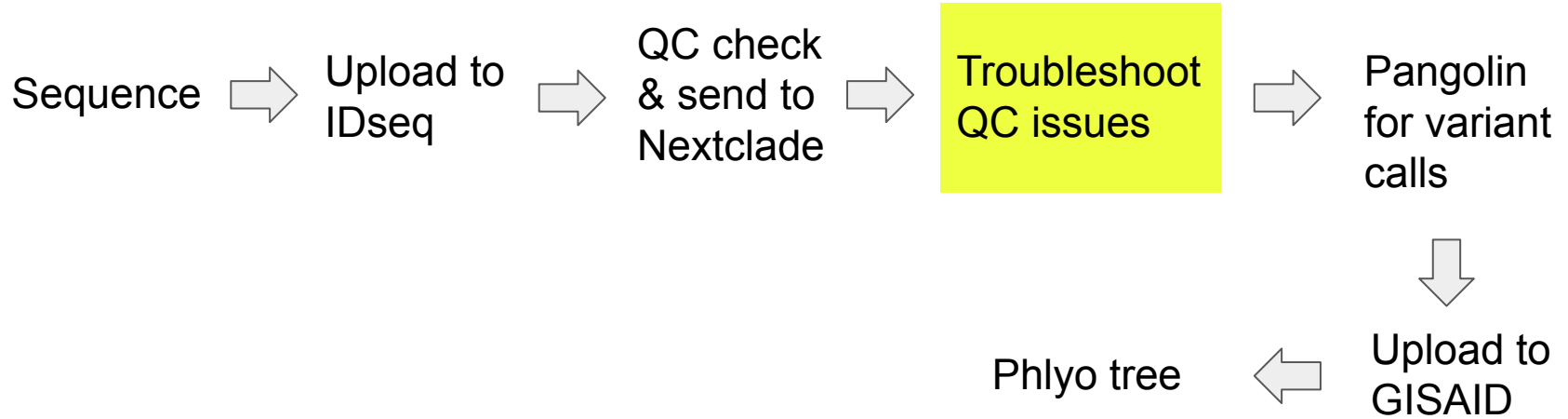
wuhan1 - AGT TGG TCC ATG ATT CGT TTC GTC TAT TCG TCT
CZI - AGT GGT CCA TGA TTC GTT TCG TTA ATT CGT CTTC

Other QC checks

Cross contamination

- Always have water controls! Negative controls also good to have
- Normal to see a handful of SARS-CoV-2 reads in controls -- but be concerned if recovering full amplicons, this is a sign of contamination.
- Plate maps -- where are the low Ct samples?

Workflow overview



N

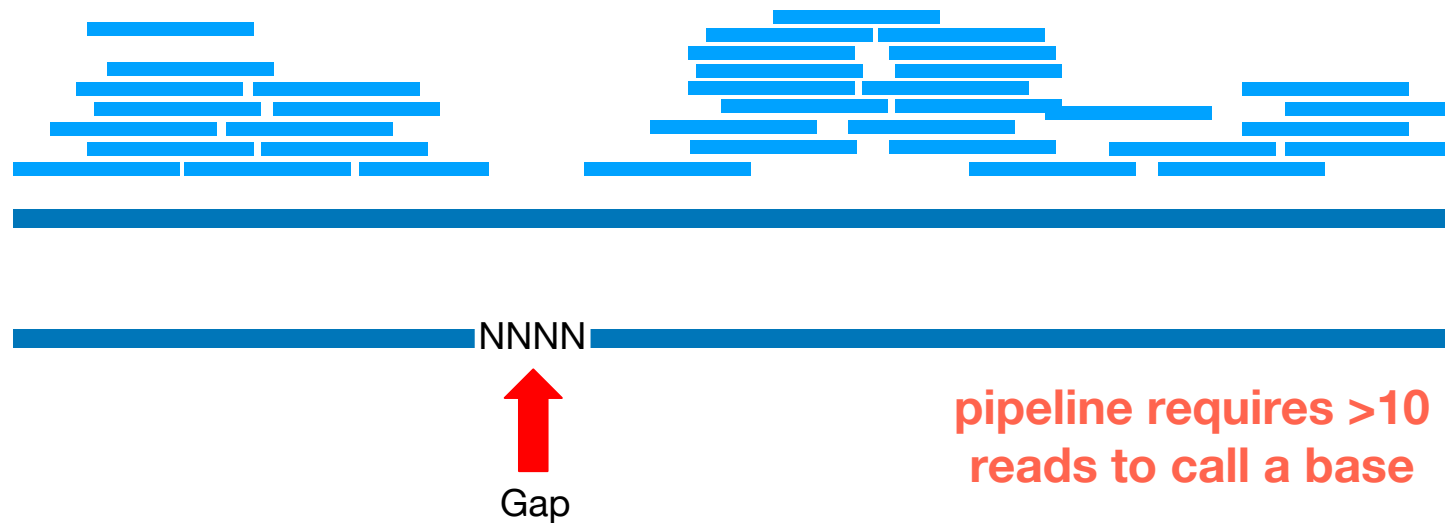
Troubleshooting too many N's

- option to resequence, but should take into account Ct value.
- can concatenate fastq files prior to IDseq upload to double the coverage
- double check sequencing metrics- was this a successful run?

Raw Reads
Aligned to
Reference

Reference
Genome

Consensus
Sequence

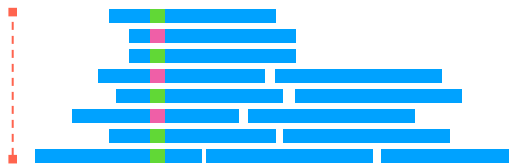


M

Troubleshooting 'mixed sites'

- Potential causes: host infected by multiple variants (rare) or contamination
- Contamination check:
 - Check plate map & barcodes used-> shared barcodes may cause bleedover during sequencing.
- Our pipeline is stringent, can check bam file to see if any bases are confidently called (ie 89% one base and 11% another).
- Make sure to pay attention to where these occur- the ends of reads tend to have lower quality bases

Raw Reads
Aligned to
Reference



Reference
Genome



Consensus
Sequence



**pipeline requires a base
to be >90% present to be
called**

P

Troubleshooting private mutations

- If there are too many private mutations- viewing the bam file can help.
- What to look for:
 - High coverage in that location all of the reads showing the same base call = good sign it that mutation is real
 - Low coverage and/or reads with different base calls = could be sign of mutations due to contamination





Troubleshooting clusters of mutations

- This usually happens after long stretches of N's

Raw Reads
Aligned to
Reference



Reference
Genome

C T C A A C

Consensus
Sequence
(100 bp)

G A G G C A

Frameshift mutations

- happen when there are there are deletions or insertions that affect the open reading frames
- Align the consensus genome back to the reference genome
- Check the open reading frames
 - You can do this in BLAST- make sure the ORFs are correct
 - If they are not, have a closer look at the alignment and check out the insertion or deletion.
- Can check bam file
- If there are frameshift mutations the CG won't be accepted to GISAID or Genbank

Descriptions Graphic Summary **Alignments** Taxonomy

Alignment view Pairwise CDS feature Re

100 sequences selected

Download GenBank Graphics

Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/hum

Sequence ID: [MT385422.1](#) Length: 29901 Number of Matches: 1

Range 1: 11 to 29862 GenBank Graphics Next Match

Score	Expect	Identities	Gaps	Strand
55127 bits(29852)	0.0	29852/29852(100%)	0/29852(0%)	Plus/Plus

Query 5 TACCTTCCAGGTAACAAACCAACCACTTCGATCTCTGTAGATCTGTTCTTAAACG 64

Sbjct 11 TACCTTCCAGGTAACAAACCAACCAACTTCGATCTCTGTAGATCTGTTCTTAAACG 70

Query 65 AACCTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACCTCACCGAGTATA 124

Sbjct 71 AACCTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACCTCACCGAGTATA 130

GenBank: [MT385422.1](#)
[GenBank](#) [FASTA](#)

Link To This View | Feedback

2 K 4 K 6 K 8 K 10 K 12 K 14 K 16 K 18 K 20 K 22 K 24 K 26 K 28 K 29.901

MT385422.1 Find: Tools Tracks Download

Sequence

Genes

QJE38304.1 ORF1ab QJE38305.1 QJE38306.1 QJE38307.1 QJE38308.1 QJE38309.1 QJE38310.1 QJE38311.1 QJE38312.1 QJE38313.1 QJE38314.1 QJE38315.1 QJE38316.1 QJE38317.1 QJE38318.1 QJE38319.1 QJE38320.1 QJE38321.1 QJE38322.1 QJE38323.1 QJE38324.1 QJE38325.1 QJE38326.1 QJE38327.1 QJE38328.1 QJE38329.1 QJE38330.1 QJE38331.1 QJE38332.1 QJE38333.1 QJE38334.1 QJE38335.1 QJE38336.1 QJE38337.1 QJE38338.1 QJE38339.1 QJE38340.1 QJE38341.1 QJE38342.1 QJE38343.1 QJE38344.1 QJE38345.1 QJE38346.1 QJE38347.1 QJE38348.1 QJE38349.1 QJE38350.1 QJE38351.1 QJE38352.1 QJE38353.1 QJE38354.1 QJE38355.1 QJE38356.1 QJE38357.1 QJE38358.1 QJE38359.1 QJE38360.1 QJE38361.1 QJE38362.1 QJE38363.1 QJE38364.1 QJE38365.1 QJE38366.1 QJE38367.1 QJE38368.1 QJE38369.1 QJE38370.1 QJE38371.1 QJE38372.1 QJE38373.1 QJE38374.1 QJE38375.1 QJE38376.1 QJE38377.1 QJE38378.1 QJE38379.1 QJE38380.1 QJE38381.1 QJE38382.1 QJE38383.1 QJE38384.1 QJE38385.1 QJE38386.1 QJE38387.1 QJE38388.1 QJE38389.1 QJE38390.1 QJE38391.1 QJE38392.1 QJE38393.1 QJE38394.1 QJE38395.1 QJE38396.1 QJE38397.1 QJE38398.1 QJE38399.1 QJE38400.1

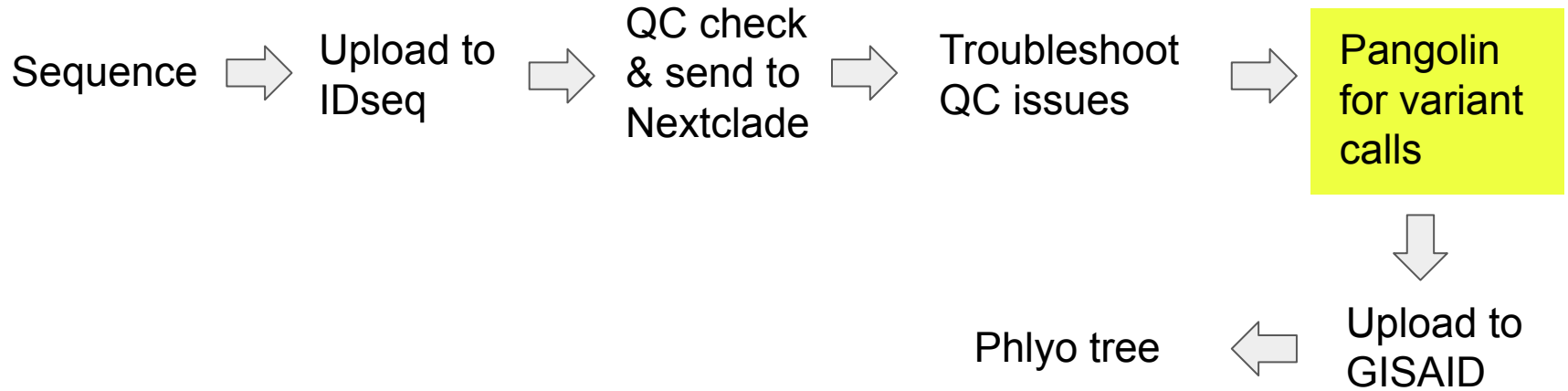


Download All

IDseq file outputs and their descriptions

File	Description	Use
consensus.fa	The consensus genome!	The consensus genome
depths.png	Coverage plots	Determine genome coverage
report.tsv	QUAST report	Quality Control
Aligned reads.bam	Initial reads that aligned to the reference genome	Can use in genome browser
ercc_stats.txt	ERCC spike in stats	Used for QC of ERCC control
no_host_1.fq.gz & no_host_2.fq.gz	Non host raw reads	Upload to SRA
Primer trimmed.bam.bai	Aligned reads with trimmed primers (companion to .bam file)	used for interrogating coverage results and ensuring quality mappings
Primer trimmed.bam	Aligned reads with trimmed primers	used for interrogating coverage results and ensuring quality mappings
stats.json	QC	Secondary QC if the coverage looks weird

Workflow overview



Download consensus genomes for variant calling and sending to public repositories

Select a Download Type

8 consensus genomes selected

Consensus Genome (consensus.fa)
Download multiple consensus genomes as separate or a single file.

Download Format:

Select format ▼

Separate Files (separate_files.csv)
Download separate files for each consensus genome, including sample collection location, collection date, and other metadata.

Consensus Genome Overview (.csv)
Consensus Genome QC metrics (e.g. % genome coverage, mapped read #, SNP #) and other summary statistics

[Start Generating Download](#)

Downloads for larger files can take multiple hours to generate.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) 8 months ago

Upload consensus genomes to pangolin



<https://cov-lineages.org/pangolin.html>



Drag and drop fasta file


Select fasta file to upload



























A pink arrow points to the 'Start analysis' button. The interface includes three buttons: 'Start analysis', 'Reset entries', and 'Upload another file'. Below the buttons is a table with two columns: 'File name' and 'Sequence name'. The table contains 8 rows of data, all with 'Consensus Genome (3).fa' in the 'File name' column. The 'Sequence name' column contains various identifiers.

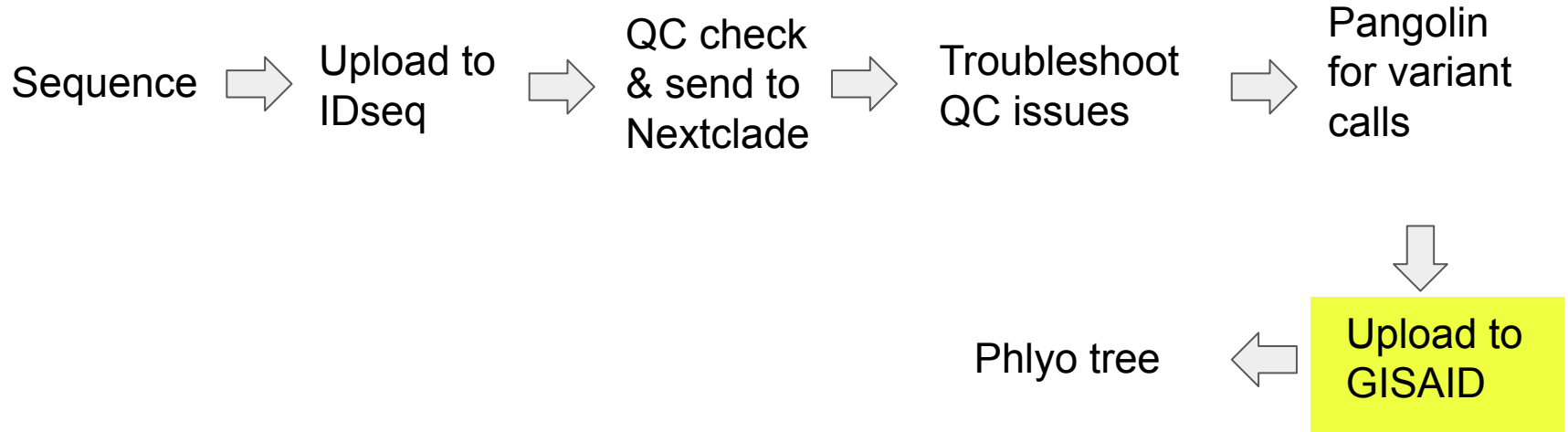
File name	Sequence name
READY FOR ANALYSIS 8 sequences	
Consensus Genome (3).fa	CASC20008_LC
Consensus Genome (3).fa	CASC20011_LC
Consensus Genome (3).fa	SRR10903402_
Consensus Genome (3).fa	SRR10971381_
Consensus Genome (3).fa	SRR10903402_
Consensus Genome (3).fa	SRR10903401_
Consensus Genome (3).fa	unknown_S1_L
Consensus Genome (3).fa	sample_sars-co

Pangolin will assign a lineage long with a probability

 [Reset entries](#) [Upload another file](#) [Help](#)

File name	Sequence name	Lineage	Assignment Conflict	↑
— ANALYSED (Click tick icon for more info) 8 sequences ↓				
✓	Consensus Genome (3).fa	CASC20008_L001 MN908947.3	B.1.1.205   	
✓	Consensus Genome (3).fa	CASC20011_L001 MN908947.3	B.1.403   	0.0
✓	Consensus Genome (3).fa	SRR10903402_44524_reads_nh_1 MN908947.3	B   	0.0
✓	Consensus Genome (3).fa	SRR10971381_44496_reads_nh MN908947.3	B   	0.0
✓	Consensus Genome (3).fa	SRR10903402_44524_reads_nh MN908947.3	B   	0.0
✓	Consensus Genome (3).fa	SRR10903401_44525_reads_nh MN908947.3	B   	0.0
✓	Consensus Genome (3).fa	unknown_S1_L001 MN908947.3	A   	0.0
✓	Consensus Genome (3).fa	sample_sars-cov-2_paired MN908947.3	B.1   	

Workflow overview



Submit fasta and metadata to GISAID

[Detailed protocol found here](#)

Upload options:

1. Single upload
2. Batch upload -> must explicitly request this function

Single Upload

Enter and upload genetic sequence and metadata, available clinical and epidemiological data, geographical as well as species-specific data. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Virus detail

Virus name*

Accession ID

Type

Passage details/history*

Sample information

Collection date*

Location*

Additional location information

Host*

Additional host information

Gender*

Patient age*

Patient status*

Specimen source

Outbreak Detail

Last vaccinated

Treatment

Submit fasta file with high quality consensus genomes

```
>hCoV-19/USA/CA-CZB-32182/2021
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCTGTTC
TCTAAACGAACCTTTAAAACTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACG
CAGTATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAA
GCAGGCTGCTTACGGTTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTTGTG
CGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTTCAACGAGAAAACACAC
GTCCAACCTAGTTTTGCCTGTTTTACAGGTTTCGCGACGTGCTCGTACGTGGCTTTGGAGAC
TCCGTGGAGGAGGCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTA
GTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTTATCAAACGT
TCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAAGCTCGAA
GGCATTAGTACGGTTCGTAGTGGTGAGACACTTGGTGTCCCTCATGTGGGCGAA
ATACCAGTGGCTTACCGCAAGGTTCTTCTTCTGTAAGAACGGTAATAAAGGAGCTGGTGGC
CATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCCAGCAGCTTGGCACTGATCCT
TATGAAGATTTTCAAGAAAACCTGGAACACTAAACATAGCAGTGGTGTACCCCGTGAACCT
ATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACCTTCTGTGGCCCT
```

```
>hCoV-19/USA/CA-CZB-32181/2021
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCTGTTCTCTAAACGA
ACTTTAAAACTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAAT
TAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAA
TTACGGTTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTTGTCCGGGTGTGA
CCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTTCAACGAGAAAACACACGTCCAACCTC
AGTTTTGCCTGTTTTACAGGTTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAG
GAGGCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGCTTAGTAGAAGTT
GAAAAAGGCGTTTTTGCCTCAACTTGAACAGCCCTATGTGTTTATCAAACGTTTCGGATGCT
CGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAAGCTCGAAGGCATTGAG
TACGGTTCGTAGTGGTGAGACACTTGGTGTCCCTTGTCCCTCATGTGGGCGAAATACCAGTG
GCTTACCGCAAGGTTCTTCTTCTGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTAC
GGCGCCGATCTAAAGTCATTTGACTTAGGCCAGCAGCTTGGCACTGATCCTTATGAAGAT
TTTTCAAGAAAACCTGGAACACTAAACATAGCAGTGGTGTACCCCGTGAACCTCATGCGTGAG
CTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACCTTCTGTGGCCCTGATGGCTAC
CCTCTTGAGTGCATTAAGACCTTCTAGCACGTGCTGGTAAAGCTTTCATGCACCTTGTCC
GAACAACCTGGACTTATTGACACTAAGAGGGGTGTACTGCTGCCGTGAACATGAGCAT
GAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTCAGACACCTTTTGAA
ATTAATTTGGCAAAGAAATTTGACATCTTCAATGGGGAATGCCAAATTTGTATTTCCC
```

Submission files

Metadata (collected during genome upload)

mandatory/optional	
Submitter	GISAID-Username
FASTA filename	the filename that contains the sequence without path
Virus name	hCoV-19/USA/CA-CZB-01/2020 (Must match name in fasta file)
Type	"betacoronavirus" (fixed)
Passage details/history	"Original" (fixed)
Collection date	
Location	North America / USA / California / Contra Costa County
Additional location information	e.g. Cruise Ship, Convention, Live animal market
Host	"Human" (fixed)
Additional host information	e.g. Patient infected while traveling in
Sampling Strategy	e.g. Sentinel surveillance (ILI), Sentinel surveillance (ARI), Sentinel surveillance (SARI), Non-sentinel-surveillance (hospital), Non-sentinel-surveillance (GP network), Longitudinal sampling on same patient(s), S gene dropout
Gender	Male, Female, or <i>unknown</i>
Patient age	e.g. 65 or 7 months, or <i>unknown</i>
Patient status	e.g. Hospitalized, Released, Live, Deceased, or <i>unknown</i>
Specimen source	Nasopharyngeal/oropharyngeal swab
Outbreak	Date, Location e.g. type of gathering, Family cluster, etc.
Last vaccinated	provide details if applicable
Treatment	Include drug name, dosage
Sequencing technology	Illumina Miseq
Assembly method	minimap2 / iVar
Coverage	e.g. 70x, 1,000x, 10,000x (average)
Originating lab	Where the clinical specimen or virus isolate was first obtained
Address	
Sample ID given by the originating laboratory	
Submitting lab	Where sequence data have been generated and submitted to GISAID
Address	
Sample ID given by the submitting laboratory	
Authors	a comma separated list of Authors with complete First followed by Last Name

[Registered Users](#)

[EpiFlu™](#)

[EpiCoV™](#)

[My profile](#)



[EpiCoV™](#)



[Search](#)



[Downloads](#)



[Upload](#)



[My Unreleased](#)

GISAID hCoV-19 Batch Upload

Upload genetic sequence as single FASTA-File and metadata, available clinical and epidemiological data, geographical as well as species-specific data as XLS or CSV. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Metadata as Excel or CSV*

max size: 5M

No file chosen

Sequences as FASTA*

max size: 32M

No file chosen

Report



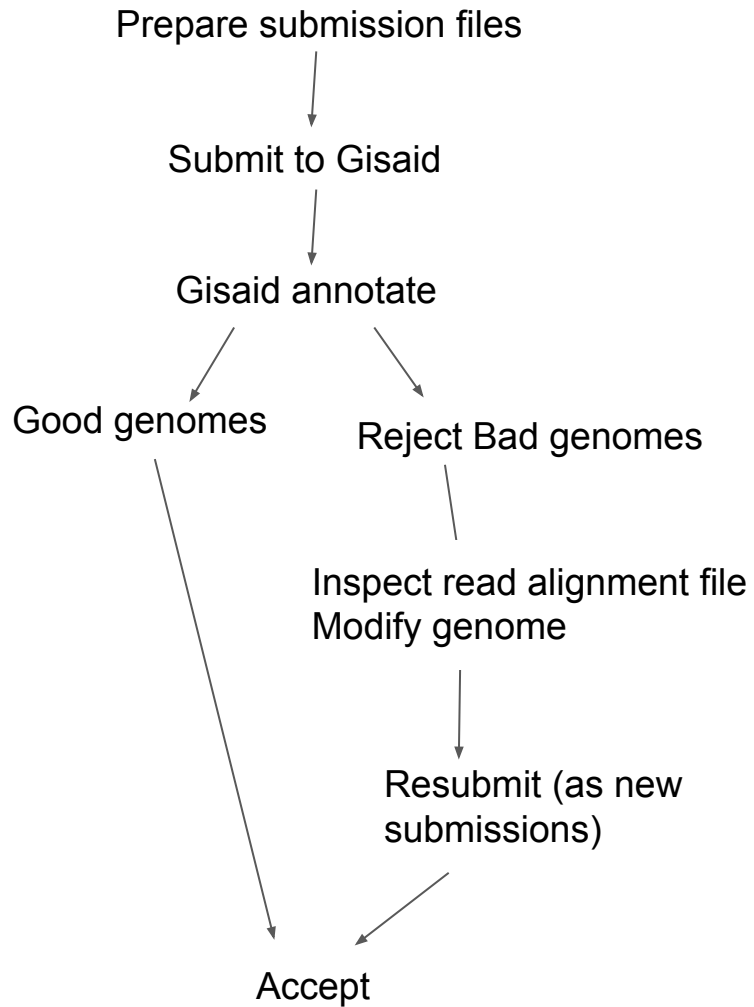
[Download Instructions and Template](#)




[Contact Curator](#)



[Check and Submit](#)



Rejected sequences will have Accession ID assigned, and resubmission of modified genomes needs to go through either a curator or the website



© 2008 - 2021 | [Terms of Use](#) | [Privacy Notice](#) | [Contact](#)

You are logged in as **Dan Lu** - [logout](#)

Registered Users
EpiFlu™
EpiCoV™
My profile

EpiCoV™
 Search
 Downloads
 Upload
 My Unreleased

My Uploads

The following submissions are currently being reviewed by a curator. Prior to release, the curator can be contacted for any changes.

<input type="checkbox"/>	edit	Virus name	Passage de	Accession ID	Collection da	Submission D		Length	Host	Location	Origin
<input type="checkbox"/>		Batch '210422_not_on_gisaid_gisaid.xls'				2021-04-22 2		659		North America / U	
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-29408/2021	Original	EPI_ISL_1664048	2021-01-05	2021-04-21		29,837	Human	North America / U	Contra
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-29412/2021	Original	EPI_ISL_1664013	2021-01-06	2021-04-21		29,848	Human	North America / U	Contra
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-29409/2021	Original	EPI_ISL_1663945	2021-01-06	2021-04-21		29,849	Human	North America / U	Contra
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-28830/2020	Original	EPI_ISL_1664028	2020-11-25	2021-04-21		29,903	Human	North America / U	Alame
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-28607/2021	Original	EPI_ISL_1663952	2021-02-09	2021-04-21		29,800	Human	North America / U	CA DF
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-29710/2021	Original	EPI_ISL_1663984	2021-03-03	2021-04-21		29,807	Human	North America / U	Santa
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-29709/2021	Original	EPI_ISL_1663925	2021-03-03	2021-04-21		29,811	Human	North America / U	Santa
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-29363/2021	Original	EPI_ISL_1664009	2021-01-04	2021-04-21		29,864	Human	North America / U	Contra
<input type="checkbox"/>		hCoV-19/USA/CA-CZB-29743/2021	Original	EPI_ISL_1664064	2021-03-05	2021-04-21		29,853	Human	North America / U	Santa

Upload to SRA



- No_host_1.fg.gz & No_host_2.fg.gz
- The raw reads with human reads filtered out
- Upload to SRA

Easily submit assembled & raw read SARS-CoV-2 data for COVID-19 response. NCBI is here to help.

GenBank

Started 2020-06-28

Submit assembled reads of SARS-CoV-2 with FASTA files and source metadata. Annotation for SARS-CoV-2 is not required.

Accessions in 1-2 working days (avg)

Submit

Sequence Read Archive (SRA)

Started 2020-06-28

Submit unassembled reads of SARS-CoV-2 with BioProject, BioSample, metadata and NGS files.

Accessions in 2 hours (avg)

Submit

Workflow overview

